Public Archives Canada | Archives publiques Canada

Canada

Machine Readable Archives BULLETIN

Vol. 1 – No. 4
Winter 1984

ISSN 0821-3658

Processing Machine Readable Data

Overview

All data files acquired by the division are processed according to internal procedures. These procedures were developed several years ago and were geared specifically to numeric files or survey data. The increased use of database management systems and the increase in the acquisition of textual and cartographic data have had an impact on particular steps in the procedures. Specific processing procedures for varying types of data are being investigated at the present time. To date it has been found that the general principles outlined for survey files can be followed for other types of data.

The processing of data files refers to the verification of the data with the record layout and the preparation of the documentation or codebook. The latter was described in the previous issue of the *Bulletin* (Vol. 1–No. 3). Processing of the data file ensures that researchers can use the files without difficulty. Discrepancies and errors in the files are uncovered and documented. The MRA does not "clean" data in the same way as do many university-based archives. Errors, inconsistencies, and unidentified codes are documented in the codebook, but are *not* corrected. As many of the data files are acquired from federal government departments and agencies, they must document the work undertaken by the department. Any "cleaning" of the data by the Archives would constitute a change to the original record and would no longer reflect the operations of the department or agency.

Processing Steps

The steps involved in the processing of the data files always appear to be straightforward. It is only when the archivist tackles a difficult file that the complications arise. The first step is the preparation of two working copies of the file. These are made using 9 track, 6250 BPI magnetic tapes with EBCDIC coding and IBM standard labels. A printout of the successful copy program is retained. If difficulties are encountered in completing the copying of the tape, it is useful to have a label dump. The information contained in the label will indicate any discrepancies in the dataset name, label, etc. A partial listing or dump of the data is obtained. This usually includes the first and last five blocks of the data. A quick check at this point will sometimes show if there is anything unusual about the data. The record layout can be specifically checked by comparing the dump with the documentation. Major variations between the logical record and the record layout can be identified. The dump is usually done in character format; however, if the data show added characters or a lack of pattern, a dump in both character and hexidecimal format may reveal the presence of data in non-character format such as packed decimal. Files for which these steps have been completed are considered to be processed to level one.

For almost all files more detailed checks are undertaken. The data are examined both for the purpose of checking the adequacy and accuracy of the documentation provided by the donor, and for the purpose of finding errors, oddities, or other anomalies. More thorough data checking is undertaken by using statistical packages such as SAS or SPSS. Discrete variables are normally checked by means of frequency counts. Out-of-range codes are identified and investigated. The investigation may require contacting the donor to obtain more information on the codes in question. Codes that cannot be explained are indicated as errors. No data are corrected, changed, or deleted.

Inter-field checks are done on variables having logical relationships with the values of other variables. For example, one field may state that a respondent is male while another contains answers to a question asked only of females. The second field should therefore contain a code for "non-applicable" or some such value. Inter-record checks are also undertaken when there is a logical relationship between the records. Once the data and record checks are completed, the documentation package is prepared. Files for which data checks have been completed are processed to level two. Two new copies of the data file are made on archival quality tapes and are then transferred for storage at two separate sites.

The above steps are oriented to survey-type files. As mentioned earlier, the general principles hold for other types of data. The division has investigated the acquisition and processing of data from database management systems, in particular System 2000. It has been possible to produce a sequentially-unloaded character version of the database. Further investigation is required in this area. Procedures for the processing of textual data are under development as the division is acquiring more automated textual records. Cartographic data are also being investigated and preliminary results indicate that many of the procedural steps may be similar; however, the verification of the information will require access to additional peripheral hardware and software packages.

As creators of data files have access to better software packages and are more concerned with the archival aspects at the time of survey design and file creation, many of the difficulties that presently occur will no longer exist. However, advancing technology and wider use of the computer in all fields will continue to provide archives with new problems in ensuring the long-term use of data files.

Ordering Machine Readable Data Files

Many requests for data files held by the division originate from outside Ottawa. In order to reduce the time required to process requests for data, the following information is required from the originator of the request. The researcher should indicate the title of the file, the principal investigator or organization that created the data file, and the date(s) of the MRDF. If data files being requested are updated yearly, the inclusive dates should be indicated. The researcher can contact the division by telephone indicating the files requested, but all requests must be followed up in writing. The majority of the division's holdings are on 9 track tape, at 6250 BPI, with standard IBM labels. The archivist receiving the request will verify that the format is appropriate. If the user requires an unlabelled tape or a different density, this should be specified. The division will try to accommodate any special requirements. Requests for copies of data files will be processed as quickly as possible. The minimum time is usually three days as the service bureau used by the division is located near Toronto. A maximum limit of two weeks is allowed. The researcher will receive a copy of the data and the documentation, along with a form indicating the file specifications. An invoice will be sent to the researcher by Financial Services, Public Archives Canada, and payment should be made to the Receiver General for Canada. The cost of the service is indicated in the fee schedule outlined on page 2. It should be pointed out that these costs are always subject to change.

Cont'd on page 2

"Ordering Machine Readable . . ."
cont'd from page 1

Machine Readable Archives Division Fee Schedule (effective April 1983)

The schedule lists the costs charged to the user when requesting one of the services provided by the division.

(1) *Copy of Codebook:* A charge of 10 cents per page (if only codebook is requested).

(2) *Tape Copying:* The charge for tape copies is according to the following formula:
20x + 20y where x = number of magnetic tapes
y = number of files.

For one file on one reel of tape the charge is: 20x + 20y or $20 + $20 = $40.

(3) *Data Extraction:* Extracts of data from large files are provided and costs are based on this formula: 20x + 20y + (z-7)
where x = number of magnetic tapes
y = number of files
z = person/hours *greater* than 7.

(4) *Statistical Analysis:* Statistical analysis of data is offered on a *limited* basis. The direct computer costs will be charged to the user. All requests for analysis must be submitted in writing, at which time an estimate of the cost will be prepared.

Update on the 1984 IASSIST Conference

Plans for the 1984 IASSIST conference to be held in Ottawa, from May 15 to 18, are being finalized. Workshops on Data Library Management, Micro-Computers, and Complex Data will be held on May 15. The workshops are intended to provide practical experience to those involved in the creation, use, management, and distribution of data.

The conference *per se* will be on May 16, 17 and 18. The overall conference theme is "Coming of Age in the Brave New World." In addition, there are three sub-themes. In terms of structure, each of the sub-themes will begin with a plenary session. The plenary will then break into three concurrent sessions that will explore specific aspects of the sub-theme. The three sub-themes are: privacy and confidentiality; advancing technology and its impact on all aspects of data collection, analysis, and distribution; and changes in the "Information Empire." The latter sub-theme will involve the changing roles and responsibilities of data archives and libraries and of organizations that create and/or collect data, as well as new types of data collection, particularly from an international perspective.

The Local Arrangements Committee has been very active in arranging activities. Three receptions will be held as well as tours of the city, Statistics Canada, and hopefully a "high tech" firm. The registration package should be distributed by mid-March. The conference location will be the Park Lane Hotel, although information on other hotels will be included in the registration package. The cost of the conference is as follows (all prices being in Canadian funds):

For IASSIST members:	
Conference and Workshops:	$ 90
Conference Only:	$ 80
Workshops Only:	$ 55
For non-members:	
Conference and Workshops:	$115
Conference Only:	$ 90
Workshops Only:	$ 65

A one-day fee of $50 for members and $55 for non-members will be charged. The fees cover the reception, an evening buffet, refreshment breaks, and tours.

For anyone involved in the creation, use, analysis, collection and distribution of machine readable data, the conference provides a useful opportunity to discuss problems, exchange ideas, and learn new ways to solve problems. More information can be obtained from the IASSIST Program Committee, c/o Machine Readable Archives Division, 395 Wellington Street, Ottawa, Ontario, K1A 0N3 — (613) 593-7772.

Notes

The Association of Public Data Users (APDU) held its annual conference in Washington, D.C., November 3 and 4. APDU, an American association, was organized in 1976 to encourage interaction among data users, producers, and distributors. One hundred and twenty attended the 8th Annual Conference. The sessions focused on the activities of the federal government in surveys and products; facilitating use of data; microcomputers; issues in data linking and matching; and the 1990 Census. Users of the census data had an opportunity to indicate which products they would like. The conference provided an occasion for users of public data to meet with producers of the data and outline their concerns and their needs. Users were also made aware of current studies being conducted by federal agencies. The personal contact with representatives from the federal agencies involved in data production was of more value than all the letters and telephone calls that users have to make to obtain information. The need for a similar association in Canada was expressed by the three Canadians who attended the meeting. It was felt that this type of meeting of public data users and producers provided invaluable information to both groups. In order to determine the support such an association would have in Canada contact Sue Gavrel, Documentation and Public Service Section, Machine Readable Archives Division, Public Archives of Canada, 395 Wellington Street, Ottawa, Ontario, K1A 0N3, (613) 593-7772 if you are interested.

«Demandes de fichiers· · ·»
suite de la page 1

Tarifs de la Division des archives ordinolingues (en vigueur depuis avril 1983)

Voici la liste des tarifs des services fournis par la division.

1) *Copie du livre de codage :* 10 cents la page (si la demande se limite à cela).

2) *Copie de la bande :* Les frais sont calculés comme suit : 20x + 20y; x = nombre de bandes magnétiques, y = nombre de fichiers.

   Ainsi, la copie d'un fichier enregistré sur une seule bande coûtera : 20x + 20y ou 20 $ + 20 $ = 40 $.

3) *Extraction de données :* Voici ce qu'il en coûte pour faire extraire des données de gros fichiers : 20x + 20y + (z-7); x = nombre de bandes magnétiques, y = nombre de fichiers, z = nombre d'heures-personnes *dépassant* 7.

4) *Analyse statistique :* Ce service est offert à *certaines conditions*. Les frais d'ordinateur sont facturés à l'usager. Toutes les demandes d'analyse doivent être soumises par écrit. On prépare une estimation des coûts pour chacune.

Conférence de 1984 de l'IASSIST

On achève actuellement les plans de la conférence de l'IASSIST qui se tiendra à Ottawa du 15 au 18 mai. Les ateliers sur la gestion des bibliothèques informatisées, les micro-ordinateurs et les données complexes auront lieu le 15 mai. Ils s'adressent à tous ceux qui créent, utilisent, gèrent et distribuent des données.

La conférence même se tiendra les 16, 17 et 18 mai. Le thème général est «Coming of Age in the Brave New World» et se subdivise en trois sous-thèmes. L'étude de chaque sous-thème débutera par une séance plénière, après quoi les participants se diviseront en trois groupes pour étudier plus à fond des aspects particuliers. Voici ces trois sous-thèmes : les renseignements personnels et confidentiels; l'évolution technologique et ses conséquences sur la collecte, l'analyse et la distribution des données; les bouleversements dans le monde de l'information. Ce dernier comprend l'évolution des rôles et des responsabilités des archives, bibliothèques et organismes qui créent et (ou) rassemblent des données, ainsi que les nouvelles méthodes de collecte des données considérées surtout d'un point de vue international.

Le comité organisateur local a fait du très bon travail. Il y aura trois réceptions ainsi que des visites guidées de la ville, de Statistique Canada et, nous l'espérons, d'une entreprise de haute technologie. Les formulaires d'inscription accompagnés de la documentation pertinente devraient être envoyés d'ici la mi-mars. La conférence aura lieu à l'hôtel Park Lane, mais vous recevrez aussi des renseignements sur d'autres hôtels. Voici les prix de la conférence (en devises canadiennes) :

Membres :	
Conférence et ateliers :	90 $
Conférence seulement :	80 $
Ateliers seulement :	55 $
Non-membres :	
Conférence et ateliers :	115 $
Conférence seulement :	90 $
Ateliers seulement :	65 $

Il sera possible d'assister à une seule journée moyennant 50 $ (55 $ pour les non-membres). Ces frais englobent la réception, un buffet, des rafraîchissements et des visites guidées.

Pour tous ceux qui ont à traiter des données ordinolingues (création, utilisation, analyse, collecte et distribution), cette conférence sera une excellente occasion de discuter de problèmes et d'échanger des idées. Pour de plus amples détails, veuillez vous adresser au Comité des programmes de l'IASSIST, a/s Division des archives ordinolingues, 395, rue Wellington, Ottawa (Ontario), K1A 0N3, au numéro (613) 593-7772.

Notes

L'Association of Public Data Users (APDU) a tenu son congrès annuel à Washington (D.C.) les 3 et 4 novembre. Il s'agit d'une association américaine fondée en 1976 en vue d'encourager les échanges entre les usagers, les producteurs et les distributeurs de données. Cent vingt personnes ont assisté au 8e congrès annuel dont les séances ont porté principalement sur les activités du gouvernement fédéral dans le domaine des enquêtes et des produits. Il a aussi été question de l'accessibilité aux données, des micro-ordinateurs, de la liaison et de la concordance des données, et du recensement de 1990. Ceux qui sont intéressés par le recensement ont eu la chance d'indiquer leur préférence quant au produit souhaité. Le congrès a permis aux usagers des données publiques de rencontrer les producteurs de ces données et de faire part de leurs préoccupations et de leurs besoins. Ces usagers ont aussi été informés des études en train d'être menées par des organismes fédéraux. Les échanges personnels avec les représentants des organismes fédéraux participant à la production des données ont été plus utiles que toutes les lettres et appels téléphoniques auxquels doivent recourir les usagers en quête d'informations. Les trois participants canadiens ont exprimé la nécessité de fonder une association similaire au Canada, car ce type de forum permet aux usagers autant qu'aux producteurs de données de se renseigner mutuellement. Pour déterminer l'appui qu'une telle association pourrait susciter au Canada, s'adresser à Sue Gavrel, Section de la documentation et du service au public, Division des archives ordinolingues, Archives publiques du Canada, 395, rue Wellington, Ottawa (Ontario), K1A 0N3, (613) 593-7772.

Archives publiques Canada
Public Archives Canada

Canada

Vol. 1 – nº 4
Hiver 1984

Archives ordinolingues BULLETIN

ISSN 0821-3658

Traitement des données ordinolingues

Vue d'ensemble

Tous les fichiers informatiques acquis par la division sont traités suivant une procédure interne établie il y a quelques années pour des fichiers numériques et des données d'enquêtes. Toutefois, l'utilisation accrue des systèmes de gestion informatiques et l'augmentation des acquisitions de données textuelles et cartographiques ont entraîné quelques changements. Nous étudions actuellement des procédures de traitement pour divers types de données. Il appert jusqu'à maintenant que les principes généraux élaborés pour les fichiers numériques s'appliquent à d'autres genres de données.

Le traitement des fichiers informatiques consiste à vérifier les données en regard du cliché d'enregistrement et à préparer le manuel de documentation ou livre de codage. Cette deuxième étape est décrite dans le dernier numéro de *Bulletin* (vol. 1, nº 3). Grâce à ce traitement, les chercheurs peuvent se servir des fichiers sans difficulté. Les incompatibilités et les erreurs contenues dans les fichiers sont décelées et consignées. Les AO ne «nettoient» pas les données comme le font de nombreuses archives universitaires. Les erreurs, les incohérences et les codes non identifiés sont enregistrés dans le livre de codage, mais *ne sont pas* corrigés. En effet, bon nombre des fichiers informatiques proviennent de ministères et d'organismes fédéraux et tout «nettoyage» des données constituerait une modification à l'original, qui ne refléterait plus alors les activités de l'établissement en question.

Étapes du traitement

Le traitement des fichiers informatiques est très simple au départ. Toutefois, certains fichiers peuvent causer des problèmes particuliers à l'archiviste. La première étape consiste à préparer deux copies de travail du fichier sur bandes magnétiques 6250 BPI, 9 pistes, codées EBCDIC et comportant des labels standards IBM. Une sortie imprimée du programme de duplication utilisé est conservée. Lorsque des difficultés surgissent pendant la duplication, il est bon d'avoir un imprimé des labels. L'information fournie par le label indiquera les incompatibilités dans le nom de l'ensemble de données, le label, etc. On obtient ainsi un imprimé ou une liste partiel(le) des données qui comprend généralement les cinq premiers et les cinq derniers blocs. Une vérification rapide à cette étape permettra parfois de déceler des irrégularités dans les données. La comparaison de l'imprimé des labels et du manuel de documentation peut faire apparaître des écarts importants entre l'enregistrement logique et le cliché d'enregistrement. On utilise habituellement des caractères pour l'imprimé des labels; cependant, si les données renferment de nouveaux caractères ou si leur structure diffère, un imprimé comportant à la fois des caractères et des hexadécimaux révélera la présence de données exprimées autrement qu'en caractères, par exemple le décimal condensé. Les fichiers qui ont subi ces étapes sont considérés comme étant traités au niveau 1.

Presque tous les fichiers sont soumis à une vérification plus complète. Les données sont examinées dans un double but : évaluer la pertinence et l'exactitude de la documentation fournie par le donateur, et déceler les erreurs, singularités ou autres anomalies. Pour ce type de vérification, on se sert de paquets-programmes comme le SAS ou le SPSS. Les variables discrètes sont généralement vérifiées par des calculs de fréquence. Les codes non pertinents sont détectés et étudiés. Il faut parfois demander des éclaircissements au donateur. Lorsque les codes sont inexplicables, on indique qu'il y a erreur. Aucune donnée n'est corrigée, changée ou supprimée.

Des vérifications interzones sont effectuées sur des variables ayant des liens logiques avec les valeurs d'autres variables. Par exemple, une zone peut indiquer qu'un répondant est de sexe masculin alors qu'une autre contient des réponses à une question réservée aux femmes. La deuxième zone devrait alors renfermer un code signifiant «sans objet» ou une valeur de ce genre. On compare aussi les enregistrements entre eux lorsqu'ils ont un lien logique. Une fois ces vérifications terminées, on prépare le manuel de documentation. Les fichiers dont les données ont été vérifiées sont traités au niveau 2. On fait deux nouvelles copies du fichier sur des bandes de qualité archivistique, qui seront entreposées à deux endroits différents.

Les étapes précédemment décrites s'appliquent à des fichiers d'enquêtes. Les principes généraux, rappelons-le, conviennent à d'autres types de données. La division a étudié la possibilité d'acquérir et de traiter des données de systèmes de gestion informatiques, notamment le Système 2000. Elle a déjà réussi à produire une version en caractères de la base de données, à vidage séquentiel, mais elle devra poursuivre ses recherches. Une procédure de traitement de données textuelles est également en préparation, car la division acquiert de plus en plus de documents textuels informatisés. On s'est également penché sur les données cartographiques et, selon les premiers résultats, une bonne partie de la procédure serait applicable; pour vérifier l'information toutefois, il faudra avoir accès à d'autres logiciels et matériels de périphériques.

Nombre des difficultés actuelles se résorberont d'elles-mêmes, car les créateurs de fichiers informatiques ont maintenant accès à de meilleurs logiciels et se préoccupent davantage des aspects archivistiques lors de la conception de l'enquête et de la constitution du fichier. Néanmoins, l'évolution technologique et l'utilisation accrue de l'ordinateur dans tous les domaines occasionneront de nouveaux problèmes aux archives qui doivent assurer l'utilisation à long terme de fichiers informatiques.

Demandes de fichiers ordinolingues

La division reçoit beaucoup de demandes de l'extérieur d'Ottawa. Pour en accélérer le traitement, le chercheur doit y inclure les renseignements suivants : titre du fichier, chercheur principal (particulier ou organisme), et date(s). Si les fichiers demandés sont mis à jour chaque année, les dates extrêmes doivent être mentionnées. Le chercheur peut présenter sa demande par téléphone, mais devra la confirmer par écrit. La plupart des documents de la division sont enregistrés sur bandes magnétiques 9 pistes, 6250 BPI, avec labels standards IBM. L'archiviste recevant la demande vérifie si le format est approprié. Si le chercheur veut une bande sans label ou de densité différente, il doit le préciser. La division essaie de satisfaire aux exigences particulières et de traiter les demandes de copies aussi vite que possible. Généralement, le délai est d'au moins trois jours, car la division fait appel à un centre de traitement situé près de Toronto, et ne dépasse pas deux semaines. Le chercheur reçoit une copie des données et de la documentation, ainsi qu'un formulaire indiquant les caractéristiques du fichier. Les Services financiers (Archives publiques du Canada) lui envoient ensuite une facture payable à l'ordre du Receveur général du Canada. La liste des tarifs figure ci-après. Signalons toutefois qu'elle est toujours sujette à changement.

suite à la page 2